EXHIBIT B

UBER_JCCP_MDL_000252298-UBER_JCCP_MDL_000252298.0014

EXHIBIT
434
DL 3/27/25
PENGAD 800-631-6989

Metadata

File Path	\EDISCO-25937_frank@uber.com\EDISCO-25937_frank@uber.com_88.zip	SEMANTIC
Filename	Sexual Misconduct Policy Revamp_1AHFtoO3W26ORVw83G24UyrHvjvqVEZP1UB0ObDFZuxs.docx	SEMANTIC
ILS Prod Date	1/2/2025	SEMANTIC
ILS Prod Vol	JCCP_MDL055	SEMANTIC
Redacted	Yes	SEMANTIC

A/C PRIVILEGED - Prepared at the direction of counsel

Sexual Misconduct Policy Revamp Q2 2017

Sytske Besemer, Criminal Justice Researcher, Trust and Safety Research Team

This document describes the process and steps we have taken (and are planning to take) to evaluate and redesign the sexual misconduct incident response policy.

Taking into account what we know about criminal justice theory and applying this to Uber's situation, I recommend a policy with:

- non-stigmatizing responses to first/second time 'offenders' (to avoid labeling and following the *risk* principle of what works in rehabilitation)
- education lightly based on cognitive behavioral interventions (obviously drivers will not participate in extensive cognitive-behavioral 'therapy', but we can still apply the same principles and focus on changing their attitudes (into the attitude that this kind of behavior is inappropriate on the Uber platform).
- crystal clear consequences of future behavior (attaining specific deterrence)

This proposed policy, as described in detail below, was developed through a combination of expertise in inappropriate/criminal behavior; quantitative analyses; and extensive communication with stakeholders across Uber. As is noted throughout the document, the sexual misconduct policy revamp necessitates an iterative process that will enable us to collect better information and update accordingly.

The sexual misconduct policy revamp process consists of four main steps and accordingly, this document describes:

A) **Categorization** (reevaluating existing categories of sexual misconduct behavior): The newly proposed categorization.

B) **Intervention** (deciding on interventions for each category of behavior): Proposed thresholds for each behavior, based on criminological theory, analyses with our own data, discussions with IRT and Safety team. The main results from the quantitative analyses show that:

- Inappropriate comments is the largest category of sexual misconduct
- The majority of drivers has only one sexual misconduct ticket (92% of those with an L4 incident, 82% of those with an L3 incident, 90% of those with an L1L2 incident). Previous tickets are not good predictors of future incidents.
- We also analyzed whether there might be other predictors of the level and number sexual misconduct incidents, looking at ratings, tenure of drivers, driving patterns, cancellations, other non-sexual misconduct safety incidents. The percentage of one star ratings is the only meaningful predictor of the total number of sexual misconduct tickets

a driver has. The percentage of trips taken on weekend nights and cancellation rate are the only variables predicting a higher level (L3L4) incident versus a lower level sexual misconduct incident.

This proposed policy is not only focused on ***deactivating*** drivers, we also want to ***educate*** drivers so they can change their behavior. Therefore, building on these results and criminological and behavioral change theories, the policy should include:

- non-stigmatizing responses to first/second time 'offenders' (to avoid labeling and following the *risk* principle of what works in rehabilitation)
- education lightly based on cognitive behavioral interventions (obviously drivers will not participate in extensive cognitive-behavioral 'therapy', but we can still apply the same principles and focus on changing their attitudes (into the attitude that this kind of behavior is inappropriate on the Uber platform).
- crystal clear consequences of future behavior (attaining specific deterrence)
- We should always reach out to the reported offender - ideally through a phone call, but if this is too expensive, we should design another solution where we can ensure that our message (including educational video) reaches the reported offender and that they acknowledge this before they can continue on the platform.

In the past, partners would be deactivated after three incidents in the same category of sexual misconduct. Instead, we propose that behaviors will be counted across categories, where behaviors of different severities have different severity scores according to this matrix.

C) **Deactivation Threshold** (deciding on thresholds for rejecting partners from the platform): We don't see a relationship between the number of trips and the number of sexual misconduct incidents. We propose a lifetime lookback period for sexual misconduct and we propose a severity score for each of the categorized behaviors. This way, all sexual misconduct behaviors are looked at together (instead of the old policy where three strikes were counted within a category of sexual misconduct behavior).

D) Evaluation of the new policy and next steps: We aim to evaluate the effectiveness of the new policy employing an experimental design. Moreover, in phase 2 of the SM policy work we will refine analyses predicting those drivers who might be involved in a second sexual misconduct incident, building further onto the results from the multivariate regression analyses.

Reevaluate the Categories and definitions

Currently (Q2 2017), the following policy exists for IRT agents where there are four basic categories of sexual misconduct (here is an archived copy). This is the policy that we will redesign. The Analyses document gives more detail about the process we went through and more details about our recommendation. Below is the proposed categorization:

Level	Category	Subcategory
Sexual Misconduct L1/L2	Non-explicit Inappropriate Remark	asking personal questions
		commenting on appearance
		flirting
		asking out / soliciting more contact
	Explicit Inappropriate Remark	
	Staring or Leering	
	Reported Consensual Sexual Activity (Driver Reports Rider-Rider)	
Sexual Assault L3/L4	Masturbation/Indecent exposure of private parts	
	Threat of Sexual Assault	
	Reported Sexual Touching	Physical Touching
		Forced Sexual Touching
	Reported Sexual Intercourse	

Intervention - Strike policy

For each of the behavioral categories, what would be the most appropriate and effective way to respond? Our policy can/should have several goals:

1. Deactivate those driver partners that are involved in serious sexual misconduct incidents (*deactivating serious offenders*).
2. Prevent drivers involved in less serious sexual misconduct incidents from showing this behavior again (this can happen either through education or through the next step: deactivation) (*specific deterrence*).

3. Deactivate drivers that are persistently involved in lower level sexual misconduct incidents (*deactivating persistent offenders which could be unintentional and/or opportunistic*).
4. Prevent drivers that have not been involved in sexual misconduct incidents from engaging in such behavior (*general deterrence*).

Types of drivers.
When we think about the policy with these goals in mind, we have 4 types of drivers:
Non-offender - Driver who has not been involved in sexual misconduct incidents.
Unintentional offender - Driver didn't realize that his behavior was offensive to a rider (could also be due to cultural differences).
Opportunistic offender - Driver could have expected that this behavior was mildly sexually inappropriate, but wasn't aware that Uber doesn't tolerate this behavior and driver poses minimal risk for serious sexual offending and is willing to change his behavior.
Serious offender - Driver either shows serious sexual assault or consistently shows (serious) sexually inappropriate behavior and is unlikely to change his behavior: deactivate (asap).

Commented [1]: +sytske@uber.com this sentence should be changed to something like "but thought they could get away with it" to be consistent with the other doc

Commented [2]: mmmm... but that sounds too narrow to me in this broader document. I don't think that all drivers that fall under this category are consciously thinking about whether or not 'they can get away' with the behavior. The main definition for this group of drivers is that they might be aware that their behavior is mildly inappropriate, but they are not aware that this behavior is not allowed on Uber. Saying they thought they could 'get away with it' sounds much more malicious, it sounds like they knew that their behavior was 'wrong'. Does that make sense?

I will discuss each of the aims and will explain where the different types of drivers are involved.

Aim 1 (Deactivate serious offenders) is focused on the serious offenders and is relatively straightforward in terms of policy: once it is clear that serious sexual misconduct has taken place the driver should be permanently rejected. The biggest challenge will be to determine what really happened and the practical challenge will be that these drivers should not be (easily) reactivated.

Aim 2 (Specific deterrence) - Specific deterrence refers to the possibility that 'sentenced' people will try to avoid the experience again. This is aimed at the unintentional as well as the opportunistic offenders. Specific deterrence only works if the individual consequences are very clear: it has to be crystal clear to drivers what the consequences are of future misconduct (Helland and Tabarrok, 2007, Drago, Galbiati, and Vertova ,2009). In our policy, we need to develop appropriate and specific responses to drivers making sure they understand the consequences of such behavior in the future (i.e. deactivation) and helping them to change their behavior in order for them to stay active as an Uber driver.

Aim 3 (Deactivating persistent offenders) - is focused on those drivers (opportunistic as well as unintentional) that are unable to change their behavior after having received several warnings / opportunities to change their behavior. In our policy, we need to develop just and effective deactivation thresholds - which we should then communicate clearly to people involved in lower level incidents so they can avoid reaching these thresholds and instead be good Uber drivers.

Aim 4 (General deterrence) is difficult to attain. In contrast to specific deterrence, general deterrence refers to the notion that threat of punishment deters most people from committing crime in the first place, even without experiencing the punishment themselves. General deterrence doesn't work: most offenders are not deterred by potential fines or prison sentences, mostly because the certainty that they'll be caught and sentenced is very low. The severity of a

crime doesn't matter, the certainty of punishment is much more important (Bottoms & von Hirsch, 2010; Cullen, Jonson, & Nagin, 2011).

I am mentioning this aim of general deterrence, because there might be a hope / aim to reach some kind of general deterrence, but I want to emphasize that we should focus our efforts instead on the previous aims: deactivating those drivers involved in serious sexual misconduct, creating crystal clear policies for drivers involved in lower level incidents to prevent them from engaging in such behaviors again *and* developing fair and effective thresholds for deactivating drivers who continue to be involved in lower level inappropriate sexual misconduct. The other lesson from general deterrence research is that we should make the certainty of punishment / consequence as high as possible: drivers should feel like there is a guardian watching over them, so if they do engage in inappropriate behaviors the likelihood of a consequence is high.

Labeling theory.
When we respond to drivers engaged in sexual misconduct, we should be careful not to label these drivers as 'sexually deviant' or 'deviant' at all, especially since we cannot always check whether this incident really happened. Classic labeling theory proposes that people act in accordance with labels attached to them by society (Becker, 1963; Lemert, 1967; Scheff, 1974). This can be a crucial factor leading from "primary deviance" (experimentation with delinquent activity) to a more persistent criminal life course. Cullen and Jonson (2014) suggest that negative labeling effects are more likely to occur with low-risk or first time offenders, who are at lower probability of continued criminality in the absence of criminal justice interference.

Labeling individuals is risky and can also create defiance: when individuals perceive a sentence as unfair they might feel excluded from a society (the Uber society) and might develop a pride that results in an increase or persistence of their deviant behavior (Sherman, 1993; 2014).

Commented [3]: +pparigi@uber.com here is a short summary of labeling with some references

What works.
In criminology, there has been a lot of research on what works in terms of rehabilitating offenders and preventing recidivism. Rehabilitation appears to be most effective when programs follow three principles of effective correctional treatment (See e.g. Andrews, 1995; Andrews & Bonta, 2010; Andrews et al., 1990; Gendreau, Little, & Goggin, 1996; Smith, Gendreau, & Swartz, 2009).

- First, risk factors for crime can be static or dynamic. Static factors such as race and sex cannot be changed, but dynamic factors such as antisocial attitudes, and association with criminal others can. The *need* principle states that interventions should focus on addressing these dynamic criminogenic needs.
- Second, treatment is effective only when it focuses on and is responsive to these criminogenic needs in a behavioral way. This is called the *responsivity* principle. The most effective treatments are cognitive-behavioral interventions focused on changing antisocial attitudes, cognitions, personality orientations related to recidivism (Cullen & Jonson, 2012).
- The third principle of effective correctional treatment is the *risk* principle, which states that interventions such as prison sentences should be given only to high-risk offenders,

because they have many criminogenic needs that can be easily targeted. Low-risk offenders, in contrast, are actually more likely to stop offending if they do not become involved in the justice system through prison sentences.

Taking into account what we know about labeling, deterrence, and rehabilitation and applying this to Uber's situation, I recommend a policy with:

- non-stigmatizing responses to first/second time 'offenders' (to avoid labeling and following the *risk* principle of what works in rehabilitation)
- education lightly based on cognitive behavioral interventions (obviously drivers will not participate in extensive cognitive-behavioral 'therapy', but we can still apply the same principles and focus on changing their attitudes (into the attitude that this kind of behavior is inappropriate on the Uber platform).
- crystal clear consequences of future behavior (attaining specific deterrence)

Theory and existing research on crime and sexual misconduct

Understanding what could be causing behavior will be helpful in designing intervention and prevention strategies. Criminal behavior, including sexual offending can be explained in many ways, including by individual factors and factors in the environment.

Non-consensual intercourse on the Uber platform can be explained by situational factors

From my reading of the more severe sexual misconduct incidents (non-consensual intercourse), situational factors play a tremendous role in these events happening. The typical scenario for a non-consensual intercourse incident is where a male driver picks up an intoxicated female rider. The female rider might actually make sexual advances towards the driver. Regardless of whether she seemed to want sexual intercourse, this behavior is not acceptable on the Uber platform: sexual intercourse is not allowed in any case and we disapprove of 'taking advantage' of the fact that the female rider is intoxicated. However, I do want to point out that the majority of these drivers would most likely not have engaged in sexual misconduct had they not been in this specific situation with this 'opportunity' for sexual misconduct. Routine Activity Theory hypothesizes that crime will take place under three necessary conditions: a motivated offender, a suitable target and the absence of a capable guardian, coming together in time and space. *This is important to keep in mind when we consider the literature on sexual offending, because sexual offending is a serious form of criminal behavior and I am not convinced that the drivers involved in incidents described above are similar to most serious sex offenders.*

Definition of sexual offending

In this sense, the definition of sexual assault on the Uber platform might be 'broader' than the definition generally used in research on sexual offending. In research on sexual offending, the participants are generally those who are convicted of a serious sexual offense and in that sense that is a more deviant group of sexual offenders. We are including a much wider range, from those drivers who make inappropriate comments to those involved in non-consensual intercourse which also ranges on a continuum from rider seemingly consenting to (physically) forcing a rider to have sex. I just want to point that out when we consider the literature on sexual

offending. Even though we might reprimand or deactivate drivers on our platform because of unwanted behaviors, this does not mean that they are sexual offenders from a legal perspective.

Escalation occurs, but not in the majority of sexual offenders

Some sexual offenders can be classified as escalators, but not all of them. Escalation of offense type was visible among 38.8% of serial rapists in a study conducted by Miethe, Olson and Mitchell (2006). Leclerc, Lussier and Deslauriers-Varin (2015) classify sexual offenders as escalators, specialists, or de-escalators, with the majority being specialists. Stermac and Hall (1989) find that escalation happened especially among younger offenders who sexually assaulted strangers and had a psychiatric history. When we consider sexual offending on the Uber platform, for some drivers involved in serious (L4) sexual misconduct, we see escalation in the sense that they were involved in lower level sexual misconduct (L1L2 inappropriate comments), but for the majority (around 90%) this L4 incident is the first incident recorded on our platform.

However, we should take into account that our data have a huge 'dark number', that part of offending that is not measured by statistics (Maguire, 2012). This means that we only see part of people's total offending behaviour, we see the 'tip of the iceberg'. People are likely to exhibit more offending behaviour that is invisible to the police and/or Uber, and many people will never appear in official statistics even though they exhibited offending and/or violent behavior. We would only know about sexual misconduct if the victim reported it to Uber or to the police. Victimization of sexual misconduct is less likely than any other serious violent crime to be reported. Even though we do not see evidence of escalation in our data, this is not conclusive evidence that it does not happen.

Considering the specific situation of sexual misconduct described above where there is an intoxicated female at night, the 'step'/'leap' from minor inappropriate comments to this specific situation is relatively small. This is why I think we should take inappropriate sexual comments, including comments about appearance, asking someone out, flirting, and leering seriously and respond to the accused offender in an appropriate way so they realize that this behavior is not tolerated on the Uber platform and we can hopefully prevent them from escalating to non-consensual intercourse.

Analyses into sexual misconduct on Uber

Next to keeping these aims and guidelines about deterrence, labeling, defiance, rehabilitation, and sexual offending in mind, we analyzed the situation of Uber drivers being involved in sexual misconduct to ensure that we create a policy that is applicable to Uber's situation. Coming back to these different types of drivers, we should aim to keep the unintentional and opportunistic drivers on the platform and educate them about behavioral expectations when driving for Uber (e.g. the cultural norms, the videos, etcetera). At the same time, we should aim to deactivate the serious offenders (also called sexual predators by Alan Berkowitz). Therefore, there were several research questions:

1) Identifying Serious Offenders

A small proportion of sexual offenders are responsible for large number of victims, and these chronic sexual offenders similarly have an early onset, a high offending frequency, and a long sexual offending career (Smallbone & Cale, 2015). How can we filter out this minority of serious sexual misconduct offenders (preferably before they are involved in a serious L3L4 sexual misconduct incident)? How do drivers involved in serious sexual misconduct (L3L4) differ from drivers involved in lower level (L1L2) sexual misconduct incidents? Can we identify who turns out to be these serious offenders? Specifically, do we see differences in:

- Ratings
- Total number of trips before the incident
- Cancellation percentages
- Total number of other safety incidents
- Driving patterns (e.g. weekends, nights, weekend nights)

Our analyses show that ***we cannot easily identify drivers who are involved in L3 and L4 incidents based on previous sexual misconduct tickets***. The only two variables that distinguished drivers involved in L3 and L4 incidents compared with drivers involved in lower level incidents were: tenure (number of trips before the incident happened was lower for higher level incidents) and driving patterns (drivers involved in L3 and L4 incidents drive more at night, on weekends, and on weekend nights). When we entered all these variables into a multivariate regression analysis, cancellation rate and the percentage of trips taken on weekend nights were the only two variables significantly related to the level of sexual misconduct.

The odds ratio for the independent variable percentage of trips taken on weekend nights is rather small, but significantly related to our outcome variable. When we look at the actual percentage for each of the groups, we see that for drivers involved in L4 incidents, the percentage of trips taken on *weekend nights* (Saturday and Sunday mornings between midnight and 6am) is on average 16.3%, for L3 13.4%, and for L1-2 10.6%.

Cancellation rate was the only other meaningful variable distinguishing drivers who were involved in level 3 and 4 incidents and drivers only involved in lower level sexual misconduct. The odds ratios are quite high, but the difference between the groups is rather small (4.5% for L1L2 and 5% for L3L4).

Since these differences are so small, IRT agents can not manually make decisions on specific cancellation rates or percentage of trips on weekend nights. Instead, for phase two of the sexual misconduct revamp, we recommend that these variables will be used to 'flag' potentially higher risk drivers. These variables can be included in future analyses where we will also include other potentially relevant variables, such as ratings by females and cancellations by drivers of male riders. Similar to what happens with dangerous driving (Escalated Dangerous Driving), once someone hits a 'threshold' (in case of dangerous driving, three tickets for that behavior), a Jira ticket is created automatically and this driver needs to be evaluated by a highly trained IRT agent who has experience with sexual misconduct incidents.

I also looked at the number of sexual misconduct tickets, disregarding the level of sexual misconduct. We see that the percentage of trips that were rated with 1 star is moderately correlated with the number of sexual misconduct incidents. In the multivariate regression

analyses, the percentage of one star ratings is the only meaningful predictor of the total number of sexual misconduct tickets a driver has. However, the differences in the percentages are small and this variable is not useful as a tool as part of the human adjudication process by IRT agents. Again, we propose that this variable could be used to 'flag' potentially riskier drivers (also combined with other 'risk' factors).

2) Identifying Early Indicators for Future Serious Offense

Are there specific types of minor incidents that could be indicators for more serious offending? This is the information we should be using to create our strike policy: e.g. do we see many different types of incidents or only specific incidents?

We do not necessarily see 'escalation' from lower level misconduct to higher level misconduct: the majority of drivers involved in higher level sexual misconduct incidents have not been involved in other incidents. I also analyzed whether different categories of behavior are related to the total number of tickets and/or L3L4 incidents, but the results are not meaningful to inform our policy for a variety of reasons:

- These analyses are based on the current categorization, and especially inappropriate comments is a very broad category with a variety of behaviors.
- The majority of people have only 1 ticket, so this data is meaningless to analyze the relationship between different types of tickets.

We can only look at incidents starting January 2016 (for Jira) and August 2016 (for Bliss). I highly recommend that we redo these analyses again in 6 or 12 months when we have more data as well as data using the new categorization to see whether we have more meaningful data to analyze escalation patterns. I also recommend taking into account which behaviors came first. Perhaps it might be helpful to do some kind of latent class analysis to see whether we can identify certain 'types' of drivers with specific behavior patterns.

Commented [4]: I have heard this and seen this written. Please note that Bliss was piloted in Oct of 2015 and fully-launched by December 2015.

3) Assessing Lookback Period

For each of these behaviors and the strike system we propose, we should consider whether previous behaviors will be considered over the lifetime of a driver or for a specific lookback period. The first question we need to ask ourselves is: when drivers drive more trips, does the prevalence of sexual misconduct incidents go up, i.e. do we see a linear relationship? If we do, we should potentially consider a lookback period REDACTED - PRIVILEGED REDACTED - PRIVILEGED Some behaviors naturally go up when drivers drive more, one can think about dangerous driving but also interpersonal incidents. This could potentially also be the case with lower level sexual misconduct incidents, such as inappropriate comments, since people might perceive a situation differently. This would be especially the case for the unintentional SM driver.

Commented [5]: REDACTED - PRIVILEGED REDACTED - PRIVILEGED

Commented [6]: REDACTED - PRIVILEGED REDACTED - PRIVILEGED

Commented [7]: REDACTED - PRIVILEGED REDACTED - PRIVILEGED

A/C PRIVILEGED - Prepared at the direction of counsel

I analyzed whether there is a correlation between the number of trips a driver has taken and the number of sexual misconduct incidents. ***We do not see any relationship at all between number of trips taken and number of sexal misconduct incidents*** (r = -0.024, 95% CI -0.032–0.016). We therefore recommend to use a lifetime lookback period, also because we really do not want such sensitive behaviors on the platform and want to respond seriously to these behaviors.

Designing the strike policy
The results of the analysis should inform our strike policy. Furthermore:

- When deciding whether someone has previous incidents, currently IRT agents only look at similar types of incidents and do not consider other types of incidents. In this situation, it is possible that drivers have two strikes for inappropriate comments, two strikes for sexism discriminatory remarks, two strikes for other types of behaviors, etcetera. We should create a policy where these different but related behaviors are considered together.

- When we try to verify whether an incident really happened, can we take into account the 'reliability' / 'trustworthiness' of a rider? For example, a rider with many rides and very few appeasements might be more credible than a rider with only a few rides (see for example) and/or a rider with many rides and plenty of appeasements. Currently, IRT agents do this intuitively, but it would be good to create a process where riders will be flagged if they have a high proportion of refunds and appeasements over total number of trips.

- If it is clear that a report is false, the contact type should be changed, so that this ticket does not count as a sexual misconduct ticket on a driver's record.

Commented [8]: Can you give an example of how we'd know that it's clear the report is false?

Commented [9]: +karna@uber.com do you happen to have an example of a ticket that is most likely false? +hrothenberg@uber.com mostly these can be seen if the rider just says 'my driver was asking personal questions, I want a refund'. If the IRT agents responds to ask some more information, they either don't respond or say 'I WANT MY MONEY BACK'. If the rider really felt treated inappropriately and they are offered money back, they generally say that they don't care about the money, but that they don't want this driver to do this again to other people. There are apparently websites telling you what to do to get a free Uber ride, see also the screenshot in this Bliss ticket (ticket not actually false but showing a picture of a list of things you can do to get a free Uber ride).: https://bliss.uberinternal.com/contacts/8bd8f91b-ddf8-4199-a9d0-78519ba4ac60

Commented [10]: If you look through a bunch of L4's on Bliss, you are bound to find one. It took maybe 30 minutes for me to hit an appeasement abuser in NYC that I use in my data integrity deck; I do not have the uuid handy.

Deactivation Thresholds for each category of behavior
Since we do not have ticket data classified in the required categories, we cannot yet meaningfully quantitatively analyze which categories are more strongly related to more serious sexual misconduct or multiple sexual misconduct tickets. Therefore, we have assigned a severity score to each of the sub-categories. Determination of the scores was based on the level of seriousness of each behavior and discussions with IRT. For example, asking personal questions can be considered inappropriate by some, but might be a conversation starter for others. Uber still needs to respond to these behaviors and tell the people involved that this behavior is undesirable when on the Uber platform, but we should not respond as severely as when someone is making an explicit sexually inappropriate remark.

After 6-12 months, when we have appropriately classified data, we will analyze the relationships between the different categories and reevaluate the weights assigned to each.

A/C PRIVILEGED - Prepared at the direction of counsel

There are two considerations for Severity Score :

1. All partners with more than 3 Sexual Misconduct tickets should automatically be flagged for manual review and classification of tickets in the new categories (e.g. checking whether tickets are legitimate and should count). Based on the same, Investigation agents will make a deactivation decision for the partner. Any driver whose Sexual-Misconduct Risk score >= 3 should be deactivated. At this stage, it is also possible that drivers will remain active if the total score of their incidents is below 3 (e.g. with 3 personal questions). Even though they might not be deactivated at this point, they will receive specific educational interventions as a result of these tickets.
2. Any conclusive support to the claims (in the manual investigation stage) will change the severity score of the ticket. This is considered important if a reporting partner submits any pictures/video to consider. Support is generally different for each category of behavior. A full list of what constitutes as support can be found here.

Based on this, following severity matrix is proposed as a framework:

Category	Sub-Category	No support	With support	Immediate Waitlist
Non-explicit Inappropriate Remark	asking personal questions	0.5	0.75	No
	commenting on appearance	0.75	1	No
	flirting	0.75	1	No
	asking out / soliciting more contact	0.75	1	No
Staring or Leering		0.75	1	No
Explicit Inappropriate Remark		1.25	1.5	No
Reported Consensual Sexual Activity (Driver Reports Rider-Rider)		1	1.5	No
Masturbation/Indecent exposure of private parts		2	3	Yes
Threat of Sexual Assault		1.5	3	Yes
Reported Sexual Touching	Physical Accidental Touching	1.5	2	Yes
	Forced Sexual Touching	2.25	3	Yes
Reported Sexual Intercourse		2.25	3	Yes

* the severity weights here are indicators and need to be finally confirmed

Summarized, L3L4 (Urgent) incidents will lead to immediate waitlisting and investigation will then determine whether someone will be permanently deactivated or reactivated. Lower level

Commented [11]: I think this is my fundamental issue with the approach: why 3? There is a great amount of qualitative research above. A decent amount of quantitative research I am still digging through, but directly shows that this is the right number. I know this is an initial starting point, but that is weak rationale. Why not 2? Why not 4?

Commented [12]: to be added, +karna@uber.com _Assigned to Deleted user_

Commented [13]: +meron@uber.com

Commented [14]: +sytske@uber.com can we make it clear somewhere in this doc that what we consider support shifts as the severity increases? I.e. in order to asking personal questions as "with support" we would need something like a video showing the person asking those questions, but for reported sexual intercourse we wouldn't require photographic support and instead could look at things like GPS location or telematics. +karna@uber.com as FYI

Commented [15]: Yes +karna@uber.com is working on a list of things that are considered support for each category of behavior.

Commented [16]: Clarify examples of support vs no support for category - due to the fact that the standard changes based on severity

Commented [17]: +sytske@uber.com REDACTED - PRIVILEGED

REDACTED - PRIVILEGED

Commented [18]: REDACTED - PRIVILEGED

REDACTED - PRIVILEGED

Commented [19]: REDACTED - PRIVILEGED

REDACTED - PRIVILEGED

Commented [20]: https://docs.google.com/document/d/1Bml8VuxGA3lNGjU7mQO8Z3vUbPfROCcYfhdT9dVzaE/edit?disco=AAAABN8ujgk

Commented [21]: +akankshu@uber.com Reported consensual activity is not currently in the new taxonomy

Commented [22]: If this is Rider-Rider, would it not be classified under "Masturbation / Indecent Exposure"

Commented [23]: +meron@uber.com referring to your comment in the JIRA taxonomy doc, here is physical accidental touching.

Commented [24]: In some documents / decks there is "Accidental", in others there is not. This somewhat highlights the judgment in the numbers - would not ...

A/C PRIVILEGED - Prepared at the direction of counsel

(L1L2) incidents will not automatically lead to waitlisting, but will require a response according to the Urgent Support Logic. The query based algorithm running automatically in the background will create a Jira ticket once someone reaches the threshold after which investigation by experienced IRT agents will determine deactivation.

REDACTED - PRIVILEGED

In phase 2 of the analyses for the sexual misconduct policy, we aim to implement the following:

3. Drivers will be assigned a Risk factor REDACTED - PRIVILEGED based on leading indicator study done by Research (based on factors such as: the percentage of one-star ratings, driving patterns, etc). At the moment, all drivers will be Normal risk until we finalize the research on high indicators in the second phase of the SM policy research.

4. Following 6-12 month of new ticket data with detailed classifications mentioned above, we will re-calculate the weights of the category based on the prediction of serious incident behaviors.

Commented [25]: +karna@uber.com +mfrangopoulos@uber.com REDACTED - PRIVILEGED REDACTED - PRIVILEGED

Commented [26]: REDACTED - PRIVILEGED REDACTED - PRIVILEGED

Commented [27]: REDACTED - PRIVILEGED REDACTED - PRIVILEGED

Commented [28]: REDACTED - PRIVILEGED REDACTED - PRIVILEGED

Commented [29]: +mfrangopoulos@uber.com REDACTED - PRIVILEGED REDACTED - PRIVILEGED

Commented [30]: REDACTED - PRIVILEGED REDACTED - PRIVILEGED

Deactivation Threshold: lookback period

I have run analyses to see whether there is a correlation between the number of trips a driver has taken and the number of sexual misconduct incidents. I have calculated this in several ways, but I do not see any relationship at all. There are, however, two considerations to keep in mind:
- I am using information on incidents from Jira starting January 2016 and Bliss starting August 2016, so the lookback period is not that far.
- perhaps it is difficult to find such a correlation as sexual misconduct incidents are rare.

Since we do not see a 'natural' increase in sexual misconduct tickets, we can consider previous sexual misconduct incidents over the lifetime (instead of for example over the last N trips or the last N months/years). This also makes sense from a safety perspective, as sexual misconduct is a type of behavior that we really want to minimize on the platform as much as we can.

Next steps: Evaluation of the new policy

Does this policy decrease recidivism of sexual misconduct on the platform?
We should aim to evaluate the new policy to see whether the new policy decreases recidivism. I propose to use an experimental design, so we can draw meaningful conclusions about the impact of the new policy. There are a few features that we can study the impact of:

- (specific deterrence) whether or not there is a crystal clear consequence of future behavior. For example, we can have two conditions after a first offense:
  a) Old policy where we send the reported offender that we don't want to see this behavior on the platform, but we don't check whether they have read the message and we don't include any educational aspects to change their behavior.
  b) New policy where we say we don't want to see this behavior on the platform and include educational content for drivers to change their behavior and crystal clear consequences in accordance with our policy.

We need to discuss how we will implement such an experimental design, as IRT agents will have to apply different 'conditions' to drivers involved in such behaviors. Karna suggested to split the driver sample by letter of their first name, so IRT agents can relatively straightforward apply condition a) or b) to a driver. We will have to find an even split in terms of first names (e.g. A-L M-Z) so both samples will be of roughly the same size.

Another option might be to instruct IRT agents in one location (e.g. Phoenix) to use the new policy, while IRT agents in another location (e.g. Chicago) continue to use the old policy. However, this is currently not possible, because Bliss tickets are automatically assigned to IRT agents, so it might be possible that a ticket first goes to an IRT agent in Phoenix, but when the rider responds it might go to an IRT agent in Chicago. L3L4 investigation IRT agents assign Jira tickets to themselves. So if we would want this design, the Bliss and Jira ticket allocation process would need to be changed, which I assume might be complicated from an operations perspective.

Without an experimental design we cannot draw any meaningful conclusions about the impact of our new policy.

Commented [31]: +hrothenberg@uber.com +karna@uber.com I would love to hear both your input here, as I can come up with an experimental design, but I don't have enough experience to know what is feasible within Uber. I wanted to tag Michael O'H here too, but he's getting married, so I purposefully did not tag him. ;) we should ask his input after he's back.

Commented [32]: I think it's fine to leave this as is in this doc as it's clear the experimental design details need to be ironed out.

This could also be a good place to work with Mike White's team to identify how best to integrate with IRT.

Do certain lower level sexual misconduct incidents distinguish L4 or persistent sexual misconduct incidents?
This policy proposed several new categories for sexual misconduct, especially for the large category of 'inappropriate comments', distinguishing between personal questions, comments on appearance, flirting, and asking someone out (ranging from innocent to more 'deviant' sexual misconduct). My hypothesis is that the more serious behaviors might predict consecutive sexual misconduct incidents and particularly more serious sexual misconduct incidents. We should evaluate whether this is indeed the case - and if so - we could consider to update the policy.
In 6-12 months time, once we have incident data based on the new categorization, we should analyze:

- Whether certain categories are more strongly related to many sexual misconduct incidents

- Whether certain categories are more strongly related to higher level sexual misconduct incidents

Based on these analyses, we can re-evaluate the severity scores for each category.

Even though we thoroughly analyzed the data that we currently have, we want to dig even deeper into the issue of sexual misconduct on the Uber platform. Specifically, we want to analyze whether we can:

- Distinguish drivers involved in higher levels
- Distinguish drivers involved in more frequent sexual misconduct
- Identify drivers likely to engage in a second or third sexual misconduct incident (using survival analyses).

Also including the following variables:

- Ratings by females, the idea being that filing a report might be cumbersome for people affected by inappropriate behavior, but giving a lower rating is an easy step. Most inappropriate behavior is targeted at females, so ratings from females should be more informative than ratings by males. As Uber does not collect gender information, we are currently working with data science (Thibault) to assign gender based on rider's first name. Since this data is on trip level, it takes a long time to gather this information and we will analyze this in the next phase, in Q3.
- Cancellations by drivers of male riders, the idea being that 'sexual predators' want female riders. Similarly to ratings, we need data on trip level for this and we will analyze this in Q3.

These and the previous results from the analyses should be used to decide whether some drivers might be flagged as potentially higher risk. If so, they might be deactivated sooner (e.g. after 2 strikes instead of 3) than those drivers with 'normal' risk.

Moreover, the analyses so far have focused only on sexual misconduct acts by drivers. In the second phase, we should also analyze sexual misconduct acts by riders, including acts of riders on riders (could be consensual for the parties involved in the acts, but not for bystanders such as drivers).

Appendix: additional reading and literature

Specific Deterrence

The idea of punishment as a specific deterrent comes from the economic model of crime in which individuals weigh costs and benefits of offending in a rational choice process. Possible punishment for a crime contributes heavily to the cost side of the equation, hypothetically reducing the probability of offending. Although this model is straightforward, in reality the costs of offending are not always clear. Offenders often do not know what sentences they would face for different crimes, and the cost of possible punishment depends on perceived unpleasantness of the experience for the individual. Although rational choice models generally predict deterrent effects of punishment, there are contingencies to this prediction, depending on perceived certainty and unpleasantness of punishment.

Specific deterrence effects may be weak because costs of offending are generally not obvious to offenders—i.e., offenders do not know what sentence they would face for certain crimes. However, Helland and Tabarrok (2007) demonstrate that imminent and extremely tangible threats of life imprisonment reduce reoffending. Helland and Tabarrok (2007) investigated the effects of the California Three Strikes and You're Out Law. They compared two groups of offenders: (1) individuals convicted of two "strikable" offenses, and (2) individuals who had been tried for a second "strikable" offense but were convicted of a non-strike-eligible offense. The imminent threat for the first group was the prospect of imprisonment for life if they committed just one more offense. The second group did not face this same threat because they had been convicted of only one "strikable" offense. The second group had a 20% higher arrest rate compared to the first group, suggesting that a specific deterrence effect was at work.
In another study, Drago, Galbiati, and Vertova (2009) investigated the effect of Italy's Collective Clemency Bill, which released more than 20,000 prisoners in 2006. Prisoners were released on the condition that individuals convicted of a different crime within 5 years of release had to serve the residual of the original sentence on top of the new sentence. Prisoners with higher residual sentences—and thus longer potential incarceration terms—had a lower subsequent recidivism rate. In both this investigation and that of Helland and Tabarrok (2007), individuals knew exactly what was going to happen if they committed another crime. Apparently, this clear information after incarceration did reduce recidivism, lending support for the idea of specific deterrence in these circumstances. We should keep this in mind for our own policy: it has to be crystal clear to drivers what the consequences are of future misconduct. In our policy, we need to develop appropriate and specific responses to drivers making sure they understand the consequences of such behavior in the future (i.e. deactivation) and helping them to change their behavior in order for them to stay active as an Uber driver.